Bayes Classifiers
• **Probabilistic Classification and Bayes’ Rule**

• **Naive Bayes Classifiers**
  - Derivation of the classification formula
  - Probability estimation and Laplace correction
  - Simple examples of naive Bayes classifiers
  - A naive Bayes classifier for the Iris data

• **Full Bayes Classifiers**
  - Derivation of the classification formula
  - Comparison to naive Bayes classifiers
  - A simple example of a full Bayes classifier
  - A full Bayes classifier for the Iris data

• **Summary**
• A classifier is an algorithm that assigns a class from a predefined set to a case or object, based on the values of descriptive attributes.

• An optimal classifier maximizes the probability of a correct class assignment.
  
  ○ Let $C$ be a class attribute with $\text{dom}(C) = \{c_1, \ldots, c_{n_C}\}$, which occur with probabilities $p_i$, $1 \leq i \leq n_C$.

  ○ Let $q_i$ be the probability with which a classifier assigns class $c_i$. ($q_i \in \{0, 1\}$ for a deterministic classifier)

  ○ The probability of a correct assignment is

\[
P(\text{correct assignment}) = \sum_{i=1}^{n_C} p_i q_i.\]

  ○ Therefore the best choice for the $q_i$ is

\[
q_i = \begin{cases} 
1, & \text{if } p_i = \max_{k=1}^{n_C} p_k, \\
0, & \text{otherwise.}
\end{cases}
\]
• Consequence: An optimal classifier should assign the most probable class.

• This argument does not change if we take descriptive attributes into account.
  ○ Let $U = \{A_1, \ldots, A_m\}$ be a set of descriptive attributes with domains $\text{dom}(A_k)$, $1 \leq k \leq m$.
  ○ Let $A_1 = a_1, \ldots, A_m = a_m$ be an instantiation of the descriptive attributes.
  ○ An optimal classifier should assign the class $c_i$ for which

$$P(C = c_i \mid A_1 = a_1, \ldots, A_m = a_m) = \max_{j=1}^{n_C} P(C = c_j \mid A_1 = a_1, \ldots, A_m = a_m)$$

• Problem: We cannot store a class (or the class probabilities) for every possible instantiation $A_1 = a_1, \ldots, A_m = a_m$ of the descriptive attributes. (The table size grows exponentially with the number of attributes.)

• Therefore: Simplifying assumptions are necessary.
Bayes’ Rule and Bayes’ Classifiers

- Bayes’ rule is a formula that can be used to “invert” conditional probabilities: Let $X$ and $Y$ be events, $P(X) > 0$. Then

$$P(Y \mid X) = \frac{P(X \mid Y) \cdot P(Y)}{P(X)}.$$ 

- Bayes’ rule follows directly from the definition of conditional probability:

$$P(Y \mid X) = \frac{P(X \cap Y)}{P(X)} \quad \text{and} \quad P(X \mid Y) = \frac{P(X \cap Y)}{P(Y)}.$$ 

- Bayes’ classifiers: Compute the class probabilities as

$$P(C = c_i \mid A_1 = a_1, \ldots, A_m = a_m) =$$

$$\frac{P(A_1 = a_1, \ldots, A_m = a_m \mid C = c_i) \cdot P(C = c_i)}{P(A_1 = a_1, \ldots, A_m = a_m)}.$$ 

- Looks unreasonable at first sight: Even more probabilities to store.
Naive Bayes Classifiers

**Naive Assumption:**
The descriptive attributes are conditionally independent given the class.

**Bayes’ Rule:**
\[
P(C = c_i \mid \omega) = \frac{P(A_1 = a_1, \ldots, A_m = a_m \mid C = c_i) \cdot P(C = c_i)}{P(A_1 = a_1, \ldots, A_m = a_m)}
\]
\[\leftarrow p_0\]
abbrev. for the normalizing constant

**Chain Rule of Probability:**
\[
P(C = c_i \mid \omega) = \frac{P(C = c_i)}{p_0} \cdot \prod_{k=1}^{m} P(A_k = a_k \mid A_1 = a_1, \ldots, A_{k-1} = a_{k-1}, C = c_i)
\]

**Conditional Independence Assumption:**
\[
P(C = c_i \mid \omega) = \frac{P(C = c_i)}{p_0} \cdot \prod_{k=1}^{m} P(A_k = a_k \mid C = c_i)
\]
Reminder: Chain Rule of Probability

• Based on the **product rule** of probability:

\[
P(A \land B) = P(A \mid B) \cdot P(B)
\]

(Multiply definition of conditional probability with \(P(B)\).)

• **Multiple application** of the product rule yields:

\[
P(A_1, \ldots, A_m) = P(A_m \mid A_1, \ldots, A_{m-1}) \cdot P(A_1, \ldots, A_{m-1})
\]
\[
= P(A_m \mid A_1, \ldots, A_{m-1})
\]
\[
\cdot P(A_{m-1} \mid A_1, \ldots, A_{m-2}) \cdot P(A_1, \ldots, A_{m-2})
\]
\[
= \vdots
\]
\[
= \prod_{k=1}^{m} P(A_k \mid A_1, \ldots, A_{k-1})
\]

• The scheme works also if there is already a condition in the original expression:

\[
P(A_1, \ldots, A_m \mid C) = \prod_{i=1}^{m} P(A_k \mid A_1, \ldots, A_{k-1}, C)
\]
Conditional Independence

- Reminder: **stochastic independence** (unconditional)

\[ P(A \land B) = P(A) \cdot P(B) \]

(Joint probability is the product of the individual probabilities.)

- Comparison to the **product rule**

\[ P(A \land B) = P(A \mid B) \cdot P(B) \]

shows that this is equivalent to

\[ P(A \mid B) = P(A) \]

- The same formulae hold conditionally, i.e.

\[ P(A \land B \mid C) = P(A \mid C) \cdot P(B \mid C) \quad \text{and} \quad P(A \mid B, C) = P(A \mid C). \]

- **Conditional independence allows us to cancel some conditions.**
(Weak) Dependence in the entire dataset: $X$ and $Y$ dependent.
No Dependence in Group 1: $X$ and $Y$ conditionally independent given Group 1.
No Dependence in Group 2: $X$ and $Y$ conditionally independent given Group 2.
• The next table shows four 3-dimensional probability distributions (one per row).

• The (in)dependencies are always w. r. t. \( A \) and \( B \).

• The condition variable is \( C \).

<table>
<thead>
<tr>
<th>marginal</th>
<th>conditional</th>
<th>( a_1 )</th>
<th>( b_1 )</th>
<th>( b_2 )</th>
<th>( a_2 )</th>
<th>( b_1 )</th>
<th>( b_2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>indep.</td>
<td>indep.</td>
<td>0.03</td>
<td>0.01</td>
<td>0.27</td>
<td>0.09</td>
<td>0.006</td>
<td>0.054</td>
</tr>
<tr>
<td>dep.</td>
<td>dep.</td>
<td>0.01</td>
<td>0.03</td>
<td>0.126</td>
<td>0.234</td>
<td>0.1275</td>
<td>0.3825</td>
</tr>
<tr>
<td>dep.</td>
<td>indep.</td>
<td>0.12</td>
<td>0.085</td>
<td>0.18</td>
<td>0.015</td>
<td>0.024</td>
<td>0.459</td>
</tr>
<tr>
<td>indep.</td>
<td>dep.</td>
<td>0.008</td>
<td>0.032</td>
<td>0.144</td>
<td>0.216</td>
<td>0.018</td>
<td>0.042</td>
</tr>
</tbody>
</table>

• All combinations are possible.
• Consequence: Manageable amount of data to store.
  Store distributions $P(C = c_i)$ and $\forall 1 \leq k \leq m : P(A_k = a_k | C = c_i)$.

• It is not necessary to compute $p_0$ explicitly, because it can be computed implicitly
  by normalizing the computed values to sum 1.

Estimation of Probabilities:

• Nominal/Symbolic Attributes

$$\hat{P}(A_k = a_k | C = c_i) = \frac{\#(A_k = a_k, C = c_i) + \gamma}{\#(C = c_i) + n_{A_k} \gamma}$$

$\gamma$ is called **Laplace correction**: Assume for every class $c_i$ some number of
  hypothetical samples for every value of $A_k$ to prevent the estimate to be 0 if
  $\#(A_k = a_k, C = c_i) = 0$.

$\gamma = 0$: Maximum likelihood estimation.

Common choices: $\gamma = 1$ or $\gamma = \frac{1}{2}$. 
Estimation of Probabilities:

- **Metric/Numeric Attributes:** Assume a normal distribution.

\[
P(A_k = a_k \mid C = c_i) = \frac{1}{\sqrt{2\pi \sigma_k(c_i)}} \exp \left( -\frac{(a_k - \mu_k(c_i))^2}{2\sigma_k^2(c_i)} \right)
\]

- Estimate of mean value

\[
\hat{\mu}_k(c_i) = \frac{1}{\#(C = c_i)} \sum_{j=1}^{\#(C=c_i)} a_k(j)
\]

- Estimate of variance

\[
\hat{\sigma}_k^2(c_i) = \frac{1}{\xi} \sum_{j=1}^{\#(C=c_i)} (a_k(j) - \hat{\mu}_k(c_i))^2
\]

\[
\xi = \#(C = c_i) : \text{Maximum likelihood estimation}
\]

\[
\xi = \#(C = c_i) - 1: \text{Unbiased estimation}
\]
Naive Bayes Classifiers: Simple Example 1

<table>
<thead>
<tr>
<th>No</th>
<th>Sex</th>
<th>Age</th>
<th>Blood pr.</th>
<th>Drug</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>male</td>
<td>20</td>
<td>normal</td>
<td>A</td>
</tr>
<tr>
<td>2</td>
<td>female</td>
<td>73</td>
<td>normal</td>
<td>B</td>
</tr>
<tr>
<td>3</td>
<td>female</td>
<td>37</td>
<td>high</td>
<td>A</td>
</tr>
<tr>
<td>4</td>
<td>male</td>
<td>33</td>
<td>low</td>
<td>B</td>
</tr>
<tr>
<td>5</td>
<td>female</td>
<td>48</td>
<td>high</td>
<td>A</td>
</tr>
<tr>
<td>6</td>
<td>male</td>
<td>29</td>
<td>normal</td>
<td>A</td>
</tr>
<tr>
<td>7</td>
<td>female</td>
<td>52</td>
<td>normal</td>
<td>B</td>
</tr>
<tr>
<td>8</td>
<td>male</td>
<td>42</td>
<td>low</td>
<td>B</td>
</tr>
<tr>
<td>9</td>
<td>male</td>
<td>61</td>
<td>normal</td>
<td>B</td>
</tr>
<tr>
<td>10</td>
<td>female</td>
<td>30</td>
<td>normal</td>
<td>A</td>
</tr>
<tr>
<td>11</td>
<td>female</td>
<td>26</td>
<td>low</td>
<td>B</td>
</tr>
<tr>
<td>12</td>
<td>male</td>
<td>54</td>
<td>high</td>
<td>A</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$P(Drug)$</th>
<th>$A$</th>
<th>$B$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.5</td>
<td>0.5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$P(Sex \mid Drug)$</th>
<th>$A$</th>
<th>$B$</th>
</tr>
</thead>
<tbody>
<tr>
<td>male</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td>female</td>
<td>0.5</td>
<td>0.5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$P(Age \mid Drug)$</th>
<th>$A$</th>
<th>$B$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mu$</td>
<td>36.3</td>
<td>47.8</td>
</tr>
<tr>
<td>$\sigma^2$</td>
<td>161.9</td>
<td>311.0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$P(Blood Pr. \mid Drug)$</th>
<th>$A$</th>
<th>$B$</th>
</tr>
</thead>
<tbody>
<tr>
<td>low</td>
<td>0</td>
<td>0.5</td>
</tr>
<tr>
<td>normal</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td>high</td>
<td>0.5</td>
<td>0</td>
</tr>
</tbody>
</table>

A simple database and estimated (conditional) probability distributions.
Naive Bayes Classifiers: Simple Example 1

\[ P(\text{Drug A} \mid \text{male, 61, normal}) \]
\[ = c_1 \cdot P(\text{Drug A}) \cdot P(\text{male} \mid \text{Drug A}) \cdot P(61 \mid \text{Drug A}) \cdot P(\text{normal} \mid \text{Drug A}) \]
\[ \approx c_1 \cdot 0.5 \cdot 0.5 \cdot 0.004787 \cdot 0.5 = c_1 \cdot 5.984 \cdot 10^{-4} = 0.219 \]

\[ P(\text{Drug B} \mid \text{male, 61, normal}) \]
\[ = c_1 \cdot P(\text{Drug B}) \cdot P(\text{male} \mid \text{Drug B}) \cdot P(61 \mid \text{Drug B}) \cdot P(\text{normal} \mid \text{Drug B}) \]
\[ \approx c_1 \cdot 0.5 \cdot 0.5 \cdot 0.017120 \cdot 0.5 = c_1 \cdot 2.140 \cdot 10^{-3} = 0.781 \]

\[ P(\text{Drug A} \mid \text{female, 30, normal}) \]
\[ = c_2 \cdot P(\text{Drug A}) \cdot P(\text{female} \mid \text{Drug A}) \cdot P(30 \mid \text{Drug A}) \cdot P(\text{normal} \mid \text{Drug A}) \]
\[ \approx c_2 \cdot 0.5 \cdot 0.5 \cdot 0.027703 \cdot 0.5 = c_2 \cdot 3.471 \cdot 10^{-3} = 0.671 \]

\[ P(\text{Drug B} \mid \text{female, 30, normal}) \]
\[ = c_2 \cdot P(\text{Drug B}) \cdot P(\text{female} \mid \text{Drug B}) \cdot P(30 \mid \text{Drug B}) \cdot P(\text{normal} \mid \text{Drug B}) \]
\[ \approx c_2 \cdot 0.5 \cdot 0.5 \cdot 0.013567 \cdot 0.5 = c_2 \cdot 1.696 \cdot 10^{-3} = 0.329 \]
• 100 data points, 2 classes

• Small squares: mean values

• Inner ellipses: one standard deviation

• Outer ellipses: two standard deviations

• Classes overlap: classification is not perfect

Naive Bayes Classifier
Naive Bayes Classifiers: Simple Example 3

- 20 data points, 2 classes
- Small squares: mean values
- Inner ellipses: one standard deviation
- Outer ellipses: two standard deviations
- Attributes are not conditionally independent given the class
• 150 data points, 3 classes
  Iris setosa     (red)
  Iris versicolor (green)
  Iris virginica  (blue)

• Shown: 2 out of 4 attributes
  sepal length
  sepal width
  petal length   (horizontal)
  petal width    (vertical)

• 6 misclassifications
  on the training data
  (with all 4 attributes)
Full Bayes Classifiers

- Restricted to metric/numeric attributes (only the class is nominal/symbolic).

- **Simplifying Assumption:**
  Each class can be described by a multivariate normal distribution.

  \[
  f(A_1 = a_1, \ldots, A_m = a_m \mid C = c_i) = \frac{1}{\sqrt{(2\pi)^m |\Sigma_i|}} \exp \left( -\frac{1}{2} (\bar{a} - \bar{\mu}_i)^\top \Sigma_i^{-1} (\bar{a} - \bar{\mu}_i) \right)
  \]

  \(\bar{\mu}_i\): mean value vector for class \(c_i\)

  \(\Sigma_i\): covariance matrix for class \(c_i\)

- Intuitively: Each class has a bell-shaped probability density.

- Naive Bayes classifiers: Covariance matrices are diagonal matrices.
  (Details about this relation are given below.)
Estimation of Probabilities:

- Estimate of mean value vector

\[ \hat{\mu}_i = \frac{1}{\#(C = c_i)} \sum_{j=1}^{\#(C = c_i)} \bar{a}(j) \]

- Estimate of covariance matrix

\[ \hat{\Sigma}_i = \frac{1}{\xi} \sum_{j=1}^{\#(C = c_i)} \left( \bar{a}(j) - \hat{\mu}_i \right) \left( \bar{a}(j) - \hat{\mu}_i \right)^\top \]

\( \xi = \#(C = c_i) \) : Maximum likelihood estimation
\( \xi = \#(C = c_i) - 1 \) : Unbiased estimation

\( \bar{x}^\top \) denotes the transpose of the vector \( \bar{x} \).

\( \bar{x} \bar{x}^\top \) is the so-called outer product or matrix product of \( \bar{x} \) with itself.
Naive Bayes classifiers for metric/numeric data are equivalent to full Bayes classifiers with diagonal covariance matrices:

\[
f(A_1 = a_1, \ldots, A_m = a_m \mid C = c_i) = \frac{1}{\sqrt{(2\pi)^m |\Sigma_i|}} \cdot \exp \left( -\frac{1}{2} (\vec{a} - \vec{\mu}_i) \top \Sigma_i^{-1} (\vec{a} - \vec{\mu}_i) \right)
\]

\[
= \frac{1}{\sqrt{(2\pi)^m \prod_{k=1}^{m} \sigma_{i,k}^2}} \cdot \exp \left( -\frac{1}{2} (\vec{a} - \vec{\mu}_i) \top \text{diag}(\sigma_{i,1}^{-2}, \ldots, \sigma_{i,m}^{-2}) (\vec{a} - \vec{\mu}_i) \right)
\]

\[
= \frac{1}{\prod_{k=1}^{m} \sqrt{2\pi \sigma_{i,k}^2}} \cdot \exp \left( -\frac{1}{2} \sum_{k=1}^{m} \frac{(a_k - \mu_{i,k})^2}{\sigma_{i,k}^2} \right)
\]

\[
= \prod_{k=1}^{m} \frac{1}{\sqrt{2\pi \sigma_{i,k}^2}} \cdot \exp \left( -\frac{(a_k - \mu_{i,k})^2}{2\sigma_{i,k}^2} \right) \equiv \prod_{k=1}^{m} f(A_k = a_k \mid C = c_i),
\]

where \( f(A_k = a_k \mid C = c_i) \) are the density functions used by a naive Bayes classifier.
Comparison of Naive and Full Bayes Classifiers

Naive Bayes Classifier

Full Bayes Classifier
Full Bayes Classifiers: Iris Data

• 150 data points, 3 classes
  Iris setosa (red)
  Iris versicolor (green)
  Iris virginica (blue)

• Shown: 2 out of 4 attributes
  sepal length
  sepal width
  petal length (horizontal)
  petal width (vertical)

• 2 misclassifications
  on the training data
  (with all 4 attributes)
Summary Bayes Classifiers

- **Probabilistic Classification**: Assign the most probable class.

- **Bayes’ Rule**: “Invert” the conditional class probabilities.

- **Naive Bayes Classifiers**
  - Simplifying Assumption: Attributes are conditionally independent given the class.
  - Can handle nominal/symbolic as well as metric/numeric attributes.

- **Full Bayes Classifiers**
  - Simplifying Assumption: Each class can be described by a multivariate normal distribution.
  - Can handle only metric/numeric attributes.